Statistical Models Exercises

September 4, 2017

1. Suppose you have an urn with blue and red balls. If N balls are selected at random with replacement (you put the ball back after you pick it), we can denote the outcomes as random variables X1,…,XN that are 1 or 0. If the proportion of red balls is p , then the distribution of each of these is Pr(Xi=1)=p.

These are also called Bernoulli trials. These random variables are independent because we replace the balls. Flipping a coin is an example of this with p=0.5.

You can show that the mean and variance are p and p(1−p) respectively. The binomial distribution gives us the distribution of the sum SN of these random variables. The probability that we see k red balls is given by: Pr(SN=k)=(Nk)pk(1−p)N−k

In R, the function dbimom gives you this result. The function pbinom gives us Pr(SN≤k).

This equation has many uses in the life sciences. We give some examples below.

The probability of conceiving a girl is 0.49. What is the probability that a family with 4 children has 2 girls and 2 boys (you can assume that the outcomes are independent)?

```
?dbinom
```

```
## starting httpd help server ... done
```

```
dbinom(2,4,0.49)
```

```
## [1] 0.3747001
```

2. What is the probability that a family with 10 children has 4 girls and 6 boys (you can assume no twins)?

```
dbinom(4,10,0.49)
```

```
## [1] 0.2130221
```

3. The genome has 3 billion bases. About 20% are C, 20% are G, 30% are T and 30% are A. Suppose you take a random interval of 20 bases, what is the probability that the GC-content (proportion of Gs or Cs) is strictly above 0.5 in this interval (you can assume independence)?

```
1 - pbinom(10,20,0.40)
```

```
## [1] 0.1275212
```

4. The following two questions are motivated by this event. The probability of winning the lottery is 1 in 175,223,510. If 189,000,000 randomly generated (with replacement) tickets are sold, what is the probability that at least one winning tickets is sold? (give your answer as a proportion not percentage)

```
p_win<-1 / 175223510
N <- 189000000
1 - dbinom(0,N,p_win) #prob of 1 or more winning tickets
```

```
## [1] 0.6599363
```

5. Using the information from the previous question, what is the probability that two or more winning tickets are sold?

```
1 - pbinom(1,N,p_win) #prob of 2 or more winning tickets
```

```
## [1] 0.293136
```

6. We can show that the binomial approximation is approximately normal when N is large and p is not too close to 0 or 1. This means that: (SN - E(SN)) / Var(SN) is approximately normal with mean 0 and SD 1. Using the results for sums of independent random variables, we can show that E(SN) = Np and Var(Sn) = Np(1 - p). The genome has 3 billion bases. About 20% are C, 20% are G, 30% are T, and 30% are A. Suppose you take a random interval of 20 bases, what is the exact probability that the GCcontent (proportion of Gs of Cs) is greater than 0.35 and smaller or equal to 0.45 in this interval? HINT: use the binomial distribution.

```
pbinom(9,20,0.4) - pbinom(7,20, 0.4)
```

```
## [1] 0.3394443
```

7. For the question above, what is the normal approximation to the probability?

```
# E(SN) = Np = 20 * 0.4
# Var(Sn) = Np(1 - p) = 20 * 0.4 * (1 - 0.4)
b <- (9 - 20*.4) / sqrt(20*.4*.6)
a <- (7 - 20*.4) / sqrt(20*.4*.6)
pnorm(b) - pnorm(a)
```

```
## [1] 0.3519231
```

8. Repeat Statistical Models Exercises #3, but using an interval of 1000 bases. What is the difference (in absolute value) between the normal approximation and the exact probability (using binomial) of the GC-content being greater than 0.35 and lesser or equal to 0.45?

```
exact_prob <- pbinom(450,1000,0.4) - pbinom(350,1000,0.4)

b <- (450 - 1000 * 0.4) / sqrt(1000 * 0.4 * 0.6)
a <- (350 - 1000 * 0.4) / sqrt(1000 * 0.4 * 0.6)
normal_approx <- pnorm(b) - pnorm(a)

abs(exact_prob - normal_approx)
```

```
## [1] 9.728752e-06
```

9. The Cs in our genomes can be methylated or unmethylated. Suppose we have a large (millions) group of cells in which a proportion p of the Cs of interest are methylated. We break up the DNA of these cells and randomly select pieces and end up with N pieces that contain the C we care about. This means that the probability of seeing k methylated Cs is binomial:

   exact = dbinom(k,N,p)

We can approximate this with the normal distribution:

```
a <- (k+0.5 - N*p)/sqrt(N*p*(1-p))
b <- (k-0.5 - N*p)/sqrt(N*p*(1-p))
approx = pnorm(a) - pnorm(b)
```

Compute the difference approx - exact for:

```
Ns <- c(5,10,50,100,500)
ps <- seq(0,1,0.25)
```

Compare the approximation and exact probability of the proportion of Cs being p, k = 1; : : : ;N 1 plotting the exact versus the approximation for each p and N combination. Study the plots and tell us which of the following is NOT true.

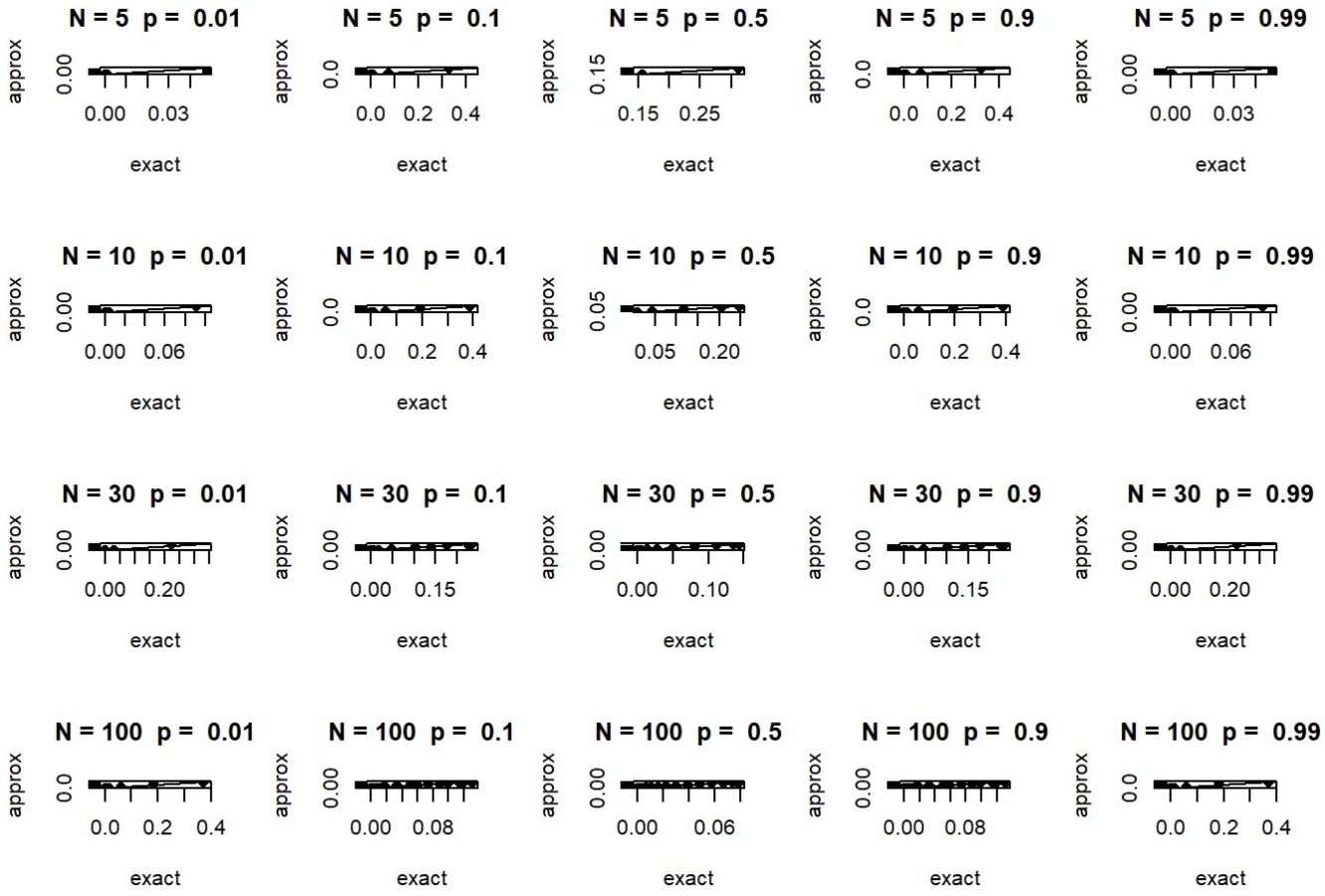
```
• A) The normal approximation works well when p is close to 0.5 even for small N = 10
• B) The normal approximation breaks down when p is close to 0 or 1 even for large N
• C) When N is 100 all approximations are spot on. (Not True)
• D) When p = 0.01 the approximation are terrible for N = 5; 10; 30 and only OK for
N = 100
```

```
Ns <- c(5,10,30,100)
ps <- c(0.01,0.10,0.5,0.9,0.99)

par(mfrow = c(4,5))
for (N in Ns){
    k <- seq(1,N-1)
    for (p in ps){
        exact = dbinom(k,N,p)

        a <- (k+0.5 - N*p)/sqrt(N*p*(1-p))
        b <- (k-0.5 - N*p)/sqrt(N*p*(1-p))
        approx = pnorm(a) - pnorm(b)

        plot(exact,approx,main=paste("N =",N," p = ",p), xlim = range(c(approx,exact))
, ylim = range(c(approx,exact)),col=1,pch=16)
        abline(0,1)
    }
}
```



10. We saw in the previous question that when p is very small, the normal approximation breaks down. If N is very large, then we can use the Poisson approximation. Earlier we computed the probability of 2 or more tickets winning the lottery when the odds of winning were 1 in 175,223,510 and 189,000,000 tickets were sold. Using the binomial we can run the code below to compute the probability of exactly two people winning to be:

    N <- 189000000 p <- 1/175223510 dbinom(2,N,p)

If we were to use the normal approximation, we would overestimate this as you can see by running this code:

```
a <- (2+0.5 - N*p)/sqrt(N*p*(1-p))
b <- (2-0.5 - N*p)/sqrt(N*p*(1-p))
pnorm(a) - pnorm(b)
```

To use the Poisson approximation here, use the rate lamda = Np representing the number of people per 20,000,000 that win the lottery. Note how much better the approximation is:

```
dpois(2,N*p)
```

In this case. it is practically the same because N is very large and Np is not 0. These are the assumptions needed for the Poisson to work. What is the Poisson approximation for more than one person winning?

```
N <- 189000000
p <- 1/175223510
dbinom(2,N,p)
```

```
## [1] 0.1978195
```

```
a <- (2+0.5 - N*p)/sqrt(N*p*(1-p))
b <- (2-0.5 - N*p)/sqrt(N*p*(1-p))
pnorm(a) - pnorm(b)
```

```
## [1] 0.2569076
```

```
dpois(2,N*p)
```

```
## [1] 0.1978195
```

```
1 - ppois(1,N*p)
```

```
## [1] 0.293136
```